

DFG-Forschergruppe "SPC" · Fakultät für Mathematik

Sergej A. Ivanov and Vadim G. Korneev

**On the preconditioning in the  
domain decomposition technique  
for the  $p$ -version finite element  
method. Part I**

**Abstract**

$P$ -version finite element method for the second order elliptic equation in an arbitrary sufficiently smooth domain is studied in the frame of  $DD$  method. Two types square reference elements are used with the products of the integrated Legendre's polynomials for the coordinate functions. There are considered the estimates for the condition numbers, preconditioning of the problems arising on subdomains and the Schur complement, the derivation of the  $DD$  preconditioner. For the result we obtain the  $DD$  preconditioner to which corresponds the generalized condition number of order  $(\log p)^2$ .

The paper consists of two parts. In part I there are given some preliminary results for  $1D$  case, condition number estimates and some inequalities for  $2D$  reference element.

The work was supported by the grant from the International Science Foundation, by the grant under the program "Universities of Russia", and by the German Academic Exchange Service (DAAD).

**Preprint-Reihe der Chemnitzer DFG-Forschergruppe  
"Scientific Parallel Computing"**

**Authors' address**

Dr. Sergej A. Ivanov / Prof. Dr. Vadim G. Korneev  
St.-Peterburg State University  
Bibliotechnaya sq. 2  
St.-Peterburg 198904  
Russia

e-mail: isa@niimm.spb.su  
korn@niimm.spb.su

## Introduction

The  $p$ -version finite element method for the second order elliptic equations for the last time has been intensively investigated theoretically and by means of numerical experiments for the possibilities of development of more efficient solution procedures. We may refer to works of I. Babuska et.al. [1] and L.F. Pavarino [2] in which  $DD$  methods for the  $p$ -version with nonoverlapping and overlapping subdomains were analyzed in this aspect. Results of series of numerical experiments and applications may be found, for instance, in J. Mandel [3], G.F. Carey and E. Barragy [4], P.F. Fisher [5], C.H. Amon [6] and others. However answers to many questions are still unclear. They are related, particularly, to the conditionality of systems of algebraic equations for the whole problem and for the problems arising on subdomains and in this connection to the choice of basis for the reference element and to the methods of solution of the problems on subdomains. It is worth to note that numerical experiments, see for instance [1, 3], showed that the most time consuming operation for  $p = 10 - 12$  even in  $2D$  case turned out to be the solution of systems of algebraic equations for subdomains. The Schur complement preconditioning in  $DD$  methods without overlapping and construction of cheap prolongation operators in different finite element spaces from the interface boundary on the whole domain demand also further study. In this paper we analyze some of the mentioned problems for the  $p$ -version at its application to the  $2D$  second order elliptic partial differential equation in an arbitrary sufficiently smooth domain.

It is assumed that one of the two reference elements or the both are used for defining the finite element space. The basis of one is  $\{\hat{L}_{i,j}(x) = \hat{L}_i(x_1)\hat{L}_j(x_2), 0 \leq i, j \leq p\}$ , where  $\hat{L}_0, \hat{L}_1$  are usual "nodal" linear functions,  $\hat{L}_k = \beta_k \tilde{L}_k$  for  $k \geq 2$  with  $\beta_k$  and  $\tilde{L}_k$  being a normalizing multiplier and the integral of the Legendre's polynomial  $L_{k-1}$  of degree  $k - 1$ . The basis of another is  $\{\hat{L}_{i,j}(x)$  for  $2 \leq i, j, (i + j) \leq p$ , for  $i = 0, 1, j = 2, 3, \dots, p$ , for  $i = 2, 3, \dots, p, j = 0, 1$  and for  $i, j = 0, 1\}$ . Special partition of the domain in quadrangles, which are curvilinear near to the boundary, and special mappings of the reference element on these quadrangles allow us to obtain the finite element method in which the boundary and the first homogeneous boundary condition are taken into account exactly. The mappings are chosen to satisfy conditions which we call the generalized conditions of quasiuniformity and which are sufficient to guarantee the same orders of convergence and condition numbers as in the case of uniform square mesh of finite elements. In the same manner the finite element  $p$ -version with the piecewise polynomial approximation of the boundary may be obtained and considered from the point of view of its numerical solution and convergence.

Condition numbers of the stiffness and mass matrices of the reference elements are estimated with the same order  $p^2$ . To obtain the first we assume that the vertices of the element are fixed. These estimates lead to estimates of condition numbers for the global finite element stiffness and mass matrices with the orders  $h^{-2}p^2 \log p$  and  $p^2$  respectively. Also some energy equivalence inequalities useful at the construction of  $DD$  preconditioners are proved.

If the bilinear form corresponds to Laplace operator then the stiffness matrix of the reference element, which is denoted via  $A_1$ , has a rather simple form. In particular, in the rows corresponding to  $\hat{L}_{i,j}$ ,  $2 \leq i, j \leq p$ , it contains only five nonzero coefficients. The conditions of the generalized quasiuniformity and simplicity of the matrix allow to conclude that it can serve as a good preconditioner for the stiffness matrices of a general curvilinear form finite elements in the case of more general elliptic equations. Considering  $A_1$  as the stiffness matrix of each element we assemble definite matrix  $\Lambda_{p,h}$ , which is a

good preconditioner for the global stiffness matrix providing the generalized condition number  $\mathcal{O}(1)$ . It turns out that the method of the nested dissection by A. George [7] is applicable for solving of the system  $\Lambda_{p,h}x = y$  with the computational cost  $\mathcal{O}(\mathcal{N}^{1.5})$ , where  $\mathcal{N}$  is the dimension of  $\Lambda_{p,h}$ . The more efficient solution techniques demand further study.

In *DD* solvers for the  $p$ -version the domains of elements or clusters of elements are usually taken for the subdomains of decomposition. For the simplicity we assume here the first. Matrix  $A_1$  as it has been noted, is a good preconditioner for the element stiffness matrices and much simpler than  $\Lambda_{p,h}$ , since it corresponds to the square. Thus, the solution of the Dirichlet problems on subdomains won't be difficult. For the Schur complement preconditioning we suggest a matrix, which provides the generalized condition number  $\mathcal{O}(\log^2 p)$ . It is represented as the product of a diagonal and a triangular matrices and so is convenient for the use in computations.

The resulting *DD*  $p$ -version conjugate gradient method demands  $\mathcal{O}(\log p \log \epsilon)$  iterations. In general situation the main time consuming operation will be the multiplication by the global stiffness matrix, which has  $\mathcal{O}(\mathcal{R}p^4)$  nonzero entries with  $\mathcal{R}$  being the number of elements. But, for instance, in the case of the Laplace equation the number of nonzero entries is  $\mathcal{O}(\mathcal{R}p^2)$ .

The article is organized as follows. In Sec.1 there are given estimates for *1D* case, which are the development of the study in [8]. The estimates for the condition numbers for the stiffness and mass matrices are derived in Sec.2. In Sec.3 the Schur complement preconditioner is obtained for the *2D* reference element and the related generalized condition number is estimated. The finite element method for the second order elliptic first boundary value problem in an arbitrary sufficiently smooth domain is described in Sec.4. There are given also the estimates of condition numbers for the global finite element matrices. The *DD* preconditioner is constructed in Sec.5. Part I contains Sec.1 and Sec.2.

Let us describe some notations used in the papers.

$I := (-1, 1)$ ,  $I^* := (0, \pi)$ ,  $\Pi := I \times I$ , also notation  $I$  with different subscripts is used for the unity matrices.

$\mathcal{P}_{p,x}$ ,  $\mathcal{P}_x^{(p)}$  are the spaces of polynomials of degree not higher than  $p$  over all variables and in each variable,  $\mathcal{P}_x^{[p]}$  is the space containing  $\mathcal{P}_{p,x}$  and polynomials of the first degree in one variable and the  $p$ -th degree in another.

$\hat{\mathcal{E}}$ ,  $\hat{\mathcal{E}}_0$  are the reference elements,  $\mathcal{E}_r$  is some finite element,  $H(\hat{\mathcal{E}})$ ,  $H(\hat{\mathcal{E}}_0)$ ,  $H(\mathcal{E}_r)$  are the spaces generated by the corresponding elements.

$$D_x^q v := \partial^{|q|} v / \partial x_1^{q_1} \partial x_2^{q_2}, \quad q = (q_1, q_2), \quad q_1, q_2 \geq 0, \quad |q| = q_1 + q_2,$$

$(\cdot, \cdot)_\Omega$ ,  $\|\cdot\|_\Omega = \|\cdot\|_{0,\Omega}$  are the scalar product and the norm in  $L^2(\Omega)$ .

$|\cdot|_{k,\Omega}$ ,  $\|\cdot\|_{k,\Omega}$  are the quasinorm and the norm in the Sobolev space  $W_2^k(\Omega)$ , i.e.

$$|v|_{k,\Omega}^2 = \sum_{|q|=k} \int_{\Omega} (D_x^q v)^2 dx, \quad \|v\|_{k,\Omega}^2 = \|v\|_{0,\Omega}^2 + \sum_{l=1}^k |v|_{l,\Omega}^2.$$

$\overset{\circ}{W}_2^1(\Omega)$  is the subspace of  $W_2^1(\Omega)$  of functions having zero traces on  $\partial\Omega$ .

$\|\cdot\|_{1/2,I}$ ,  ${}_0\|\cdot\|_{1/2,I}$ , are the norms in the space  $W_2^{1/2}(I)$  and the subspace  ${}_0W_2^{1/2}(I) \subset W_2^{1/2}(I)$  of functions having zero values at  $x = \pm 1$ . These norms for  $I^{**} = (a, b)$  are defined by expressions

$$\begin{aligned}\|v\|_{1/2, I^{**}}^2 &:= \int_a^b \int_a^b \left( \frac{u(x) - u(y)}{x - y} \right)^2 dx dy, \\ \|v\|_{1/2, I^{**}}^2 &:= \|v\|_{1/2, I^{**}}^2 + 2 \int_a^b \frac{u^2(x)}{x - a} dx + 2 \int_a^b \frac{u^2(x)}{x - b} dx.\end{aligned}$$

The norm  $\|\cdot\|_{1/2, \gamma_i}$ , where  $\gamma_i$  is the side of  $\Pi$  is defined analogously with  $\|\cdot\|_{1/2, I}$ . For instance, for the side  $\gamma_i$ , which is on the line  $x_1 = x_0$ , we have

$$\|v\|_{1/2, \gamma_i}^2 := \int_0^\pi \int_0^\pi \left( \frac{u(x_0, t) - u(x_0, \tau)}{t - \tau} \right)^2 dt d\tau.$$

Also we need the norm

$$\|u\|_{1/2, \partial\Pi}^2 = \sum_{i=1}^4 \|v\|_{1/2, \gamma_i}^2 + \sum_{i=1}^4 \int_0^\pi \frac{u_{j(i)}(t) - u_{l(i)}(t)}{|t|} dt,$$

where the  $u_{j(i)}$  denotes the restriction of  $u$  on side  $\gamma_{j(i)}$ ,  $t$  is the distance to  $v_i$ ,  $v_i$  is the vertex of  $\Pi$  such that the vertex is common for  $\gamma_{j(i)}$  and  $\gamma_{l(i)}$ .

$\mathcal{A}^+$  is the pseudoinverse to matrix  $\mathcal{A}$ .

$\lambda(\mathcal{A})$ ,  $\lambda_{\min}(\mathcal{A})$ ,  $\lambda_{\max}(\mathcal{A})$  are an eigenvalue, the minimal nonzero and the maximal eigenvalues of symmetric nonnegative matrix  $\mathcal{A}$ .

Signs  $\prec$ ,  $\succ$ ,  $\asymp$  assume one sided and two sided inequalities, which are true with some omitted absolute constants.

## 1 One dimensional finite element with the integrated Legendre polynomials as shape functions

Let us consider the system of functions on  $I$

$$\tilde{\mathcal{M}} = \{\tilde{L}_i | \tilde{L}_0 = L_0, \tilde{L}_1 = L_1, \tilde{L}_j = \int_{-1}^x L_{j-1}(s) ds, j = 2, 3, \dots, 2N\},$$

containing two first Legendre polynomials, i.e. constant  $L_0 \equiv 1$  and linear  $L_1 = x$  functions, and the first integrals of Legendre polynomials. As it is known

$$\|L_i\|^2 = 2/(2i + 1), \quad \tilde{L}_j(x) = \frac{1}{2j - 1} [L_j(x) - L_{j-2}(x)], \quad (1.1)$$

The number of functions in  $\tilde{\mathcal{M}}$  is adopted equal to  $2N + 1$  only for convenience, the considerations and results are not changed when  $i = 0, 1, \dots, 2N + 1$ .

In the following it is convenient to use another normalization and to divide sometimes  $\tilde{\mathcal{M}}$  in two subsystems, corresponding to odd and even numbers. Let us set

$$\begin{aligned}\mathcal{M} &= \{\hat{L}_i, i = 0, 1, \dots, 2N\}, \quad \bar{\mathcal{M}} = \{\bar{L}_i, i = 0, 1, \dots, 2N\}, \\ \mathcal{M} &= \bar{\mathcal{M}} = \bar{\mathcal{M}}_+ \cup \bar{\mathcal{M}}_-, \\ \bar{\mathcal{M}}_+ &= \{\bar{L}_i = \hat{L}_{2i}, i = 0, 1, \dots, N\}, \\ \bar{\mathcal{M}}_- &= \{\bar{L}_i = \hat{L}_{2(i-N)-1}, i = N + 1, \dots, 2N\},\end{aligned} \quad (1.2)$$

with  $\hat{L}_i = \tilde{L}_i / \|\tilde{L}_i\|$ ,  $\|\hat{L}_i\| = 1$ , i.e.

$$\begin{aligned} \hat{L}_0 &= \frac{L_0}{\sqrt{2}}, \quad \hat{L}_1 = \sqrt{\frac{3}{2}}\tilde{L}_1, \quad \hat{L}_j = \beta_j\tilde{L}_j = \gamma_j[L_j - L_{j-2}], \quad j \geq 2, \\ \beta_j &= \frac{1}{2}\sqrt{(2j-3)(2j-1)(2j+1)}, \quad \gamma_j = \frac{1}{2}\sqrt{\frac{(2j-3)(2j+1)}{2j-1}}, \end{aligned} \tag{1.3}$$

Systems  $\mathcal{M}$ ,  $\bar{\mathcal{M}}$  differ only in the ordering of their elements.

Expression

$$f = \sum_{i=0}^{2N} \bar{b}_i \bar{L}_i$$

and quadratic forms  $(f, f)_I$ ,  $(f', f')_I$  generate matrices

$$(f, f)_I = \langle \bar{K}_0 \bar{b}, \bar{b} \rangle, \quad (f', f')_I = \langle \bar{K}_1 \bar{b}, \bar{b} \rangle,$$

where  $\langle \cdot, \cdot \rangle$  is the scalar product in  $\mathbb{R}^{2N+1}$  and  $\bar{b} = \{\bar{b}_i\}$  is the vector of the coefficients  $\bar{b}_i$ . It is easy to note, that

$$\bar{K}_0 = \begin{pmatrix} \bar{K}_{0,+} & 0 \\ 0 & \bar{K}_{0,-} \end{pmatrix}, \quad \bar{K}_1 = \begin{pmatrix} \bar{K}_{1,+} & 0 \\ 0 & \bar{K}_{1,-} \end{pmatrix}$$

and matrices  $\bar{K}_{1,+}$ ,  $\bar{K}_{1,-}$  are diagonal, matrices  $\bar{K}_{0,+}$ ,  $\bar{K}_{0,-}$  are tridiagonal. Via  $K_0$ ,  $K_1$  we denote the same matrix  $\bar{K}_0$ ,  $\bar{K}_1$  but with the ordering of rows and columns, corresponding to representation

$$f = \sum_{i=0}^{2N} \hat{b}_i \hat{L}_i.$$

The aim of this paragraph is

**Lemma 1.1** *The nonzero eigenvalues of  $K_0$ ,  $K_1$  satisfy relations*

$$\begin{aligned} \lambda_{\min}(K_1) &= \sqrt{6}, \quad \lambda_{\max}(K_1) = (4N-3)(4N+1)/2 \\ \lambda_{\min}(K_0) &\asymp 1/N^2, \quad \lambda_{\max}(K_0) \asymp 1. \end{aligned} \tag{1.4}$$

*Proof.* These estimates are evident except for  $\lambda_{\min}(K_0)$ . Indeed, taking into account (1.1)–(1.3), we see, that

$$K_1 = \text{diag} \left[ 0, \sqrt{6}, \dots, \frac{(2i-3)(2i+1)}{2}, \dots, \frac{(4N-3)(4N+1)}{2} \right], \tag{1.5}$$

and, thus, relations (1.4) for  $K_1$  are right.

As it is shown below matrix  $\bar{K}_0$  has diagonal predominance for any  $N < \infty$ . Matrix  $K_0$  can be represented in the form  $K_0 = \mathcal{D}^{-1} A_0 \mathcal{D}^{-1}$  with the diagonal matrix  $\mathcal{D}$ , which is the square root of the diagonal of  $A_0$ . Matrix  $A_0$  is the “mass” matrix in the basis  $\{L_0, L_1, (L_2 - L_0), \dots, (L_i - L_{i-2}), \dots, (L_N - L_{N-2})\}$ :

$$A_0 = \begin{pmatrix} 2 & 0 & -2 & & & & \\ 0 & \frac{2}{3} & 0 & -\frac{2}{3} & & & 0 \\ -2 & 0 & (2 + \frac{2}{5}) & 0 & \frac{2}{5} & & \\ 0 & \dots & \dots & \dots & \dots & \dots & \\ & -\frac{2}{2i-3} & 0 & \frac{2}{2i-3} + \frac{2}{2i+1} & 0 & -\frac{2}{2i+1} & \\ & & \dots & \dots & \dots & \dots & \\ 0 & & \dots & \dots & & & \\ & & & \frac{2}{4N-3} & 0 & \frac{2}{4N-3} & \frac{2}{4N+1} \end{pmatrix}.$$

Consequently for the result we have

$$K_0 = \begin{pmatrix} 1 & 0 & -\sqrt{5/6} & & & & \\ & 1 & 0 & -\sqrt{7/10} & 0 & & \\ & & \cdot & \cdot & \cdot & & \\ & & & 1 & 0 & -c_i & \\ \text{SYM} & & & & \cdot & \cdot & \cdot \\ & & & & & 1 & 0 \\ & & & & & & 1 \end{pmatrix},$$

$$c_i = \frac{1}{2} \left( \frac{(2i+3)(2i-5)}{(2i-3)(2i+1)} \right)^{1/2} = \frac{1}{2} \left( 1 - \frac{12}{4i^2 - 4i - 3} \right)^{1/2}.$$

Easy calculations lead to inequality

$$\frac{1}{2} - c_i \geq \frac{1}{2(i^2 + 1)}, \quad i \geq 4,$$

from which and considerations of the first four rows of  $K_0$  we have

$$1 - c_i - c_{i-2} \succ 1/(i^2 + 1), \quad i = 0, 1, \dots, 2N, \quad \lambda_{\min}(K_0) \succ 1/N^2. \quad (1.6)$$

It is easy to note that the matrix, which is denoted below by  $\Delta_0$  and is obtained from  $K_0$  by making in each row the diagonal coefficients equal to the sum of modules of the non-diagonal coefficients, retains positive definiteness. Since it is meaningful for calculations we shall also estimate the boundaries of the spectrum for  $\Delta_0$ . It is sufficient to consider only  $K_{0,+}$  and the corresponding part  $\Delta_{0,+}$  of  $\Delta_0$  because for  $K_{0,-}$  considerations are the same. Thus, let us put  $\bar{K}_{0,+} = \Delta_{0,+} + \mathcal{D}_{0,+}$  where  $\mathcal{D}_{0,+}$  is the diagonal matrix,  $\Delta_{0,+}$  has the form

$$\Delta_{0,+} = \begin{pmatrix} 1/h_0 & -1/h_0 & & & & & \\ -1/h_0 & (1/h_0 + 1/h_1) & -1/h_1 & & & & 0 \\ & \dots & \dots & \dots & & & \\ & & -1/h_{i-1} & (1/h_{i-1} + 1/h_i) & -1/h_i & & \\ 0 & & & & \dots & \dots & \\ & & & & & -1/h_{N-1} & (1/h_{N-1} + 1/h_N) \end{pmatrix}$$

and  $h_i = 1/c_{2i}$ . As  $\bar{K}_{0,+}$  has diagonal predominance, so all diagonal elements of  $\mathcal{D}_{0,+}$  are positive. The above expression for  $\Delta_{0,+}$  is the same as for the finite element matrix, generated in the case of linear finite elements with the nodes

$$x^{(0)} = 0, \quad x^{(i)} = h_0 + h_1 + \dots + h_{i-1}, \quad x^{(N+1)} = b$$

by bilinear form  $(u', v')_{(0,b)}$  at the first boundary condition at  $x = b$ . That means positive definiteness of  $\Delta_{0,+}$  at any  $N < \infty$ . Consequently,  $\lambda_{\min}(\bar{K}_{0,+}) \geq \lambda_{\min}(\Delta_{0,+}) + \lambda_{\min}(\mathcal{D}_{0,+}) > 0$  if  $N < \infty$ .

For the estimation of  $\lambda_{\min}(\Delta_{0,+})$  one may use the relation

$$\lambda_{\min}(\Delta_{0,+}) = \lambda_{\max}^{-1}(\Delta_{0,+}^{-1})$$

since  $\Delta_{0,+}^{-1}$  may be written explicitly, see [9], on the basis of the above mentioned analogy with the finite element matrix. Namely we have

$$\Delta_{0,+}^{-1} = \{\kappa^{(i,j)}\}, \quad \kappa^{(i,j)} = K(x^{(i)}, x^{(j)}), \quad i, j = 0, 1, \dots, N$$

$$K(x, t) = \{b - x, x \geq t, b - t, t \leq x\}.$$

For  $\lambda_{\max}(\Delta_{0,+}^{-1})$  we can use the Frobenius estimate by the maximal sum of the modulus of coefficients in each row. Taking into account that the sum of  $K(x, x^{(j)})$  over  $j = 0, 1, \dots, N$  is equivalent to the quadrature of trapeziums, which at  $x = x^{(i)}$  is equal to the corresponding integral, we get

$$\begin{aligned} \lambda_{\max}(\Delta_{0,+}^{-1}) &= \max_{0 \leq i \leq N} \sum_j K(x^{(i)}, x^{(j)}) \leq \\ &2 \max \left[ \frac{1}{h_0}, \max_{1 \leq k \leq N} \frac{1}{h_{k-1} + h_k} \right] \int_a^b K(x, t) dt \leq \\ &\leq \max \left[ \frac{1}{h_0}, \max_{1 \leq k \leq N} \frac{1}{h_{k-1} + h_k} \right] b^2 \leq 8N^2. \end{aligned} \quad (1.7)$$

In (1.7) it is also taken into account that numbers  $c_i$  are uniformly bounded from below by  $\sqrt{33/175}$  and from above by  $1/2$ . Thus minimal eigenvalues of  $\Delta_{0,+}$  and  $\mathcal{D}_{0,+}$  are estimated from below with the same order.

Let us show now, that estimate (1.6) for  $\lambda_{\min}(K_0)$  is exact in the order. For vector of the form

$$b_+ = (0, 0, \dots, b_l, b_{l+1}, \dots, b_N)^T, \quad l = \text{entire} \frac{N}{2}, \quad b_+ \in \mathbb{R}^{N+1}, \quad (1.8)$$

we have

$$b_+^T \mathcal{D}_{0,+} b_+ \leq \frac{4}{N^2} b_+^T \cdot b_+, \quad (1.9)$$

On the other hand for the symmetrical positive definite  $\mathcal{D}_{0,+}$ ,  $\Delta_{0,+}$  it is true, that

$$b_+^T \bar{K}_{0,+} b_+ = b_+^T (\mathcal{D}_{0,+} + \Delta_{0,+}) b_+ \geq b_+^T \left( \frac{4}{N^2} I + \Delta_{0,+} \right) b_+. \quad (1.10)$$

Now we put in  $b_+$  from (1.8)  $b_j = 1$ ,  $j = l, l+1, \dots, N$ , and turn again to the explicit form of  $\Delta_{0,+}^{-1}$ . For such  $b_+$  we obtain

$$b_+^T \Delta_{0,+}^{-1} b_+ \geq \sum_{l \leq i, j \leq N} K(x^{(i)}, x^{(j)}) \geq 2 \left[ \min_{l < i < N} \frac{1}{h_{i-1} + h_i} \right]^2 \int_{b/2}^b \int_{b/2}^b K(x, t) dt dx.$$

The integral is equal to  $N^3/48$  and  $c_i = 0.5 - \mathcal{O}(N^2)$ ,  $l \leq i \leq N$ . Therefore for sufficiently large  $N$  it is true the estimate  $b_+^T \Delta_{0,+}^{-1} b_+ \geq cN^3$  and since  $b_+^T b_+ = N - l \leq 0.5N + 1$ , then

$$b_+^T \Delta_{0,+}^{-1} b_+ \geq cN^2 b_+^T \cdot b_+$$



and  $c$  is an absolute constant. From this inequality and formulas (1.9), (1.10) it follows, that  $\lambda_{\min}(\bar{K}_{0,+}) \leq cN^2$  and therefore estimate (1.6) for  $\lambda_{\min}(K_0)$  is exact in order. We have proved (1.4) for  $\lambda_{\min}(K_0)$ .

The last estimate (1.4) is evidently valid since the sum of the modulus of coefficients in each row is less than 4 and the diagonal coefficients are equal 1. Lemma has been proved.

In the finite element method the functions

$$\hat{L}_0 = \frac{1}{2}(1+x), \quad \hat{L}_1 = \frac{1}{2}(1-x), \quad (1.11)$$

are used instead of the first two functions (1.3). In the following we understand  $\mathcal{M}$ ,  $\bar{\mathcal{M}}$  as such bases. It is evident that the estimates for the bounds of the nonzero spectrums of  $\bar{K}_0, \bar{K}_1$  are retained in the order.

According to the fact of the uniform boundness of numbers  $c_i$  from below and from above matrix  $\Delta_{0,+}$  is equivalent in the spectrum to matrix

$$\Delta_{(1)} = \begin{pmatrix} 1 & -1 & & & & & \\ -1 & 2 & -1 & & & & 0 \\ & \cdots & \cdots & \cdots & & & \\ & & & -1 & 2 & -1 & \\ 0 & & & & \cdots & \cdots & \cdots \\ & & & & & -1 & 2 \end{pmatrix}.$$

The same is true for  $\Delta_{0,-}$  under agreement that the same notation  $\Delta_{(1)}$  is used for matrices (1.12) of different dimensions. In other words

$$\begin{aligned} b_+^T \Delta_{(1)} b_+ &\prec b_+^T \Delta_{0,+} b_+ \prec b_+^T \Delta_{(1)} b_+, \quad b_+ \in \mathbb{R}^{N+1}, \\ b_-^T \Delta_{(1)} b_- &\prec b_-^T \Delta_{0,-} b_- \prec b_-^T \Delta_{(1)} b_-, \quad b_- \in \mathbb{R}^N. \end{aligned} \quad (1.12)$$

## 2 Square element with the hierarchical shape functions produced by the integrated Legendre's polynomials

By the reference element  $\hat{\mathcal{E}} = \hat{\mathcal{E}}\{\Pi, \hat{L}_{i,j}, 0 \leq i, j \leq p\}$ ,  $p = 2N$ , we shall mean square  $\Pi := I \times I$  with the specified on it system  $\mathcal{M}_{\Pi} = \{\hat{L}_{i,j}, 0 \leq i, j \leq p\}$  of functions

$$\hat{L}_{i,j}(x) = \hat{L}_i(x_1)\hat{L}_j(x_2). \quad (2.1)$$

The space spanned over them is the space  $H(\hat{\mathcal{E}}) = \mathcal{P}_x^{(p)}$  of polynomials

$$\hat{u}(x) = \sum u_{i,j} \hat{L}_{i,j}(x), \quad 0 \leq i, j \leq p, \quad (2.2)$$

containing all polynomials of degree not higher than  $p$  in each variable  $x_1, x_2$ . We shall consider matrices  $A_0, A_1$  of bilinear forms

$$(\hat{u}, \hat{v})_{\Pi} = u_p^T A_0 v_p, \quad a_{\Omega}(\hat{u}, \hat{v}) = \int_{\Pi} \nabla \hat{u} \cdot \nabla \hat{v} dx = u_p^T A_1 v_p, \quad u_p, v_p \leftrightarrow \hat{u}, \hat{v}, \quad (2.3)$$

where relations  $u_p \leftrightarrow \hat{u}, v_p \leftrightarrow \hat{v}, \dots$  are understood as isomorphism between polynomials  $\hat{u}, \hat{v} \in \mathcal{P}_x^{(p)}$  and vectors  $u_p = \{u_{i,j}\}, v_p = \{v_{i,j}\} \dots$  from  $\mathbb{R}^{(p+1)^2}$  of coefficients of their

representations in the basis  $\{\hat{L}_{i,j}\}$ , see (2.2). Matrices  $A_0, A_1$  are the mass and stiffness matrices of the reference element, while energy is defined by the Dirichlet integral.

For the following system  $\mathcal{M}_\Pi$  it is convenient to subdivide into subsystems  $\mathcal{M}_I, \mathcal{M}_{II}, \mathcal{M}_{III}$  containing, correspondingly, the so called internal, side and vertex shape functions:

$$\begin{aligned}\mathcal{M}_I &= \{\hat{L}_{i,j}, 2 \leq i, j \leq p\}, \\ \mathcal{M}_{II} &= \{\hat{L}_{i,j}, i = 0, 1; j = 2, 3, \dots, p \text{ or } j = 2, 3, \dots, p; j = 0, 1\}, \\ \mathcal{M}_{III} &= \{\hat{L}_{i,j}, i = 0, 1 \text{ and } j = 0, 1\}.\end{aligned}$$

By the internal stiffness matrix we shall call matrix  $A_{1,0}$  generated by the set  $\mathcal{M}_I$ . Evidently it corresponds to the first boundary condition on  $\partial\Pi$  since all functions of  $\mathcal{M}_I$  are equal zero on  $\partial\Pi$  and there is no such functions in  $\mathcal{M}_{II}, \mathcal{M}_{III}$ .

**Lemma 2.1** *There are valid estimates*

$$\lambda_{\min}(A_0) \asymp N^{-4}, \quad \lambda_{\max}(A_0) \asymp 1, \quad \lambda_{\min}(A_{1,0}) \asymp 1, \quad \lambda_{\max}(A_{1,0}) \asymp N^2. \quad (2.4)$$

Proof. Matrices  $A_0, A_{1,0}$  are represented by means of the Kronecker products

$$A_0 = K_0 \times K_0, \quad A_{1,0} = K_{1,0} \times K_{0,0} + K_{0,0} \times K_{1,0},$$

where  $K_{1,0}, K_{0,0}$  are matrices, which are obtained from  $K_0, K_1$  by crossing out the first two rows and columns and  $K_0, K_1$  are matrices described in Sec.1. According to the properties of the Kronecker product  $\{\lambda_{m,n}(A \times B)\} = \{\lambda_m(A)\} \times \{\lambda_n(B)\}$  and consequently estimates for  $\lambda_{\min}$  from below and for  $\lambda_{\max}$  from above directly follow from (1.2). However, estimate  $\lambda_{\min}(A_{1,0}) \asymp N^{-2}$  obtained in such a way is rough and to obtain the estimates given in (2.4) it is necessary to use another ways. Let us use the representation  $\tilde{K}_{0,0} = \Delta_{0,0} + D_{0,0}$ , which is analogous to the representation  $K_{0,+} = \Delta_{0,+} + D_{0,+}$ .

Then

$$\begin{aligned}\lambda_{\min}(A_{1,0}) &= \lambda_{\min}(K_{1,0} \times (\Delta_{0,0} + D_{0,0}) + (\Delta_{0,0} + D_{0,0}) \times K_{1,0}) \\ &\geq \lambda_{\min}(K_{1,0} \times D_{0,0}) + D_{0,0} \times K_{1,0}.\end{aligned}$$

Matrices  $K_{1,0}, D_{0,0}$  are diagonal and their elements of the  $i$ -th row have the orders  $i^2, (i^2 + 1)^{-1}$ . Consequently,

$$\lambda_{\min}(A_{1,0}) \asymp \inf_{i,j} \left( \frac{i^2}{j^2 + 1} + \frac{j^2}{i^2 + 1} \right) \asymp 1, \quad 2 \leq i, j \leq 2N, \quad (2.5)$$

and estimate for  $\lambda_{\min}(A_{1,0})$  of lemma is also valid.

In order to obtain the estimate from above let us consider in  $\mathbb{R}^{(2N-1)^2}$  vectors  $w_p$  of the form  $w_p = a \times b$ , where  $a = \{a_i\} \in \mathbb{R}^{(2N-1)}$  and  $b = \{b_i\} \in \mathbb{R}^{(2N-1)}$ . We can write

$$\inf_{u_p \in \mathbb{R}^{(2N-1)^2}} \frac{u_p^T A_{1,0} u_p}{u_p^T u_p} \leq \inf_{w_p} \frac{w_p^T A_{1,0} w_p}{w_p^T w_p} = 2 \inf_{a,b} \frac{a^T K_{1,0} a}{a^T a} \frac{b^T K_{0,0} b}{b^T b} \leq 2 \inf_a \frac{a^T K_{1,0} a}{a^T a} \sup_b \frac{b^T K_{0,0} b}{b^T b}$$

from where and from Lemma 1.1 the estimate for  $\lambda_{\min}(A_{1,0})$  from above follows.

Matrix  $K_{0,0}$  contains unities on the diagonal and  $\hat{K}_{1,0}$  has diagonal coefficients of order  $N^2$ , what makes valid the estimates for  $\lambda_{\max}$  from below. According to the mentioned property of the Kronecker product  $\lambda_{\min}(A_{0,0}) = (\lambda_{\min}(K_{0,0}))^2$ , and estimates for  $\lambda_{\min}(K_{0,0})$  are known. Lemma has been proved.



Submatrix  $(p^2 - 1) \times (p^2 - 1)$  of  $K_{1,0} \times K_0$  with entries corresponding to  $i = 2, 3, \dots, p$ ;  $j = 0, 1, \dots, p$  and  $(p^2 - 1) \times (p^2 - 1)$  submatrix of  $K_0 \times K_{1,0}$  with entries corresponding to  $i = 1, 2, \dots, p$ ;  $j = 2, 3, \dots, p$  are three-diagonal. The diagonal predominance of these two submatrices are estimated exactly in the same manner as for  $K_{1,0} \times K_{0,0}$  and  $K_{0,0} \times K_{1,0}$ . Denoting via  $d_{i,j}(A)$  the diagonal predominance in row “ $i, j$ ” of some matrix  $A$  we obtain in the result that

$$d_{i,j}(\bar{A}_{1,1}) \succ \delta_{o,i} \delta_{1,i} \frac{i^2}{j^2 + 1} + \delta_{o,j} \delta_{1,j} \frac{j^2}{i^2 + 1} \quad (2.6)$$

where  $\delta_{k,l}$  is the Kronecker delta and  $i, j$  are not equal to 0 or 1 simultaneously. Lemma has been proved.

Let us give several additional inequalities related to the preconditioning of  $A_1$

**Lemma 2.3** *Let  $\hat{u} \in H(\hat{\mathcal{E}})$  be presented in the form  $\hat{u} = \hat{u}_I + \hat{u}_{II} + \hat{u}_{III}$ , where  $\hat{u}_L \in \text{span } \mathcal{M}_L$ ,  $L = I, II, III$  and*

$$\hat{a}(v, w) = \int_{\Pi} \nabla v \nabla w \, dx$$

Then

$$\begin{aligned} \frac{c_1}{1 + \log p} [\hat{a}(\hat{u}_I + \hat{u}_{II}, \hat{u}_I + \hat{u}_{II}) + \hat{a}(\hat{u}_{III}, \hat{u}_{III})] &\leq \hat{a}(\hat{u}, \hat{u}) \\ &\leq 2[\hat{a}(\hat{u}_I + \hat{u}_{II}, \hat{u}_I + \hat{u}_{II}) + \hat{a}(\hat{u}_{III}, \hat{u}_{III})], \end{aligned} \quad (2.7)$$

$$\frac{c_1}{p + \log p} \sum_{L=I,II,III} \hat{a}(\hat{u}_L, \hat{u}_L) \leq 3 \sum_{L=I,II,III} \hat{a}(\hat{u}_L, \hat{u}_L), \quad (2.8)$$

with an absolute constant  $c_1, c_2$ .

Proof. The right inequalities are Cauchy ones. The left inequality (2.7) indeed has been proved in [1] and is based on the following result, which is needed below.

**Theorem 2.1** (*I. Babuska et al. [1]*). *For any polynomial  $u(x) \in \mathcal{P}_{p,x}$  and  $x \in I$*

$$|u(x)| \leq (1 + \log^{\frac{1}{2}} p) \|u\|_{\frac{1}{2}, I} \quad (2.9)$$

with absolute constant. This estimate cannot be asymptotically improved, i.e. there is a constant  $\bar{c}$  and for each  $p \geq 2$  there exists  $v_p \in \mathcal{P}_{p,x}$  such that  $\|v_p\|_{\frac{1}{2}, I} \leq \bar{c}$  and  $|v_p(-1)| \geq \log^{\frac{1}{2}} p$ .

Now we have

$$\begin{aligned} \hat{a}(\hat{u}_I + \hat{u}_{II}, \hat{u}_I + \hat{u}_{II}) &= \hat{a}(\hat{u} - \hat{u}_{III}, \hat{u} - \hat{u}_{III}) \leq 2(\hat{a}(\hat{u}, \hat{u}) + \hat{a}(\hat{u}_{III}, \hat{u}_{III})), \\ \hat{a}(\hat{u}_I + \hat{u}_{II}, \hat{u}_I + \hat{u}_{II}) + \hat{a}(\hat{u}_{III}, \hat{u}_{III}) &\leq 2\hat{a}(\hat{u}, \hat{u}) + 3\hat{a}(\hat{u}_{III}, \hat{u}_{III}). \end{aligned} \quad (2.10)$$

Since  $\hat{u}_{III}$  is bilinear its maximum is at one of the vertices  $x = (\pm 1, \pm 1)$  of  $\Pi$  and at these points  $\hat{u}_{III} = \hat{u}(x)$ . Consequently  $\hat{a}(\hat{u}_{III}, \hat{u}_{III}) \leq c_4 \max |u(x)|^2$  over  $x = (\pm 1, \pm 1)$  and application of (2.9) and the trace theorem gives

$$\hat{a}(\hat{u}_{III}, \hat{u}_{III}) \leq c_4 c_3 (1 + \log p) \|\hat{u}\|_{\frac{1}{2}, \gamma_i} \leq c_5 c_4 c_3 (1 + \log p) \|\hat{u}\|_{1, \Pi},$$

where  $\gamma_j, j = 1, 2, 3, 4$ , are the sides of  $\Pi$  and above  $\gamma_i$  is the side adjacent to the vertex, at which  $\hat{u}_{III}$  achieves maximum. Now the Bramble–Hilbert lemma type arguments allow to write

$$a(\hat{u}_{III}, \hat{u}_{III}) \leq c(1 + \log p) |\hat{u}|_{1,\Pi}^2 \leq c(1 + \log p) a(\hat{u}, \hat{u}) \quad (2.11)$$

with an absolute constant  $c$ . Combining with (2.10) we obtain the left part of (2.7).

In order to obtain the left inequality (2.8) we use the Markov type inequality

$$|u|_{1,I}^2 \leq cp \|u\|_{1/2,I}^2, \quad \text{for any } u \in \mathcal{P}_{p,x}, \quad (2.12)$$

which is obtained in the following way. Setting  $x = \cos \phi$  we represent  $\tilde{u}(\phi) := u(x)$ ,  $\phi \in I^* := (0, \pi)$  by the sum

$$\tilde{u}(\phi) = \sum_{k=0}^p b_k \cos k\phi.$$

We also can write

$$|\tilde{u}|_{1,I}^2 \leq |\tilde{u}|_{1,I^*}^2 = \pi/2 \sum_{k=0}^p (1 + k^2) b_k^2 \leq \pi p \sum_{k=0}^p \sqrt{(1 + k^2)} b_k^2 \leq cp \|\tilde{u}\|_{1/2,I^*}^2 \leq cc_1 p \|u\|_{1/2,I}^2,$$

where in the last step we used the equivalence of the norms  $\|u\|_{1/2,I}$  and  $\|\tilde{u}\|_{1/2,I^*}$ , which has been established in [1].

It is easy to see that  $\hat{u}|_{\partial\Omega} = \hat{u}_{II} + \hat{u}_{III}$  and

$$a(\hat{u}_{II} + \hat{u}_{III}, \hat{u}_{II} + \hat{u}_{III}) \asymp \sum_{i=1}^4 \|\hat{u}\|_{1,\gamma_i}^2.$$

Thus applying Cauchy inequality, (2.11), (2.12), and the trace theorem we have

$$\begin{aligned} a(\hat{u}_{II}, \hat{u}_{II}) &\leq (a(\hat{u}_{III}, \hat{u}_{III}) + c \sum_{i=1}^4 \|\hat{u}\|_{1,\gamma_i}^2) \\ &\leq c((1 + \log p) \|\hat{u}\|_{1,\Pi}^2 + p \sum_{i=1}^4 \|\hat{u}\|_{1/2,\gamma_i}^2) \leq c(\log p + p) \|\hat{u}\|_{1,\Pi}^2. \end{aligned}$$

The use of Bramble–Hilbert lemma type arguments gives

$$a(\hat{u}_{II}, \hat{u}_{II}) \leq c(\log p + p) a(\hat{u}, \hat{u}). \quad (2.13)$$

Besides, (2.11), (2.13) allow to obtain

$$a(\hat{u}_I, \hat{u}_I) \leq 3(a(\hat{u}, \hat{u}) + a(\hat{u}_{II}, \hat{u}_{II}) + a(\hat{u}_{III}, \hat{u}_{III})) \leq c(\log p + p) a(\hat{u}, \hat{u}). \quad (2.14)$$

From (2.11), (2.13), and (2.14) it follows left inequality (2.8). Lemma has been proved.

**Remark 2.2** *In the  $p$ -version it is more efficient sometimes to use the reference element with polynomial spaces, which are the minimal polynomial spaces containing for a given  $p$  space  $\mathcal{P}_{p,x}$  and allowing to satisfy the compatibility conditions. For the chosen type of the coordinate functions this assumes the use of the reference element  $\hat{\mathcal{E}}_0 = \hat{\mathcal{E}}_0\{\Pi, \hat{L}_{i,j} \in \mathcal{M}_p\}$  with*

$$\mathcal{M}_p = \mathcal{M}_{I,p} \cup \mathcal{M}_{II} \cup \mathcal{M}_{III}, \quad \mathcal{M}_{I,p} := \{\hat{L}_{i,j}, 2 \leq i, j; (i + j) \leq p\}.$$

*Thus, for a fixed  $p$  set  $\mathcal{M}_p$  differs from  $\mathcal{M}_\Pi$  only in the subset  $\mathcal{M}_{I,p} \neq \mathcal{M}_I$  of the internal functions but the sets of the boundary functions and the vertex functions coincide. The*

space of polynomials  $H(\hat{\mathcal{E}}_0) = \text{span } \mathcal{M}_p$ , which will be denoted also by  $\mathcal{P}_x^{[p]}$ , corresponds to element  $\hat{\mathcal{E}}_0$ . The reasons for  $\hat{\mathcal{E}}_0$  to be more effective in comparison with  $\hat{\mathcal{E}}$ , especially if we enlarge  $p$  in the course of solution of one problem are known from the literature. The obtained results are retained for element  $\hat{\mathcal{E}}_0$ . Lemma 2.1 is retained in particular due to inclusion

$$\mathcal{M}_{\Pi}^{(\text{entire } p/2)} \subset \mathcal{M}_p \subset \mathcal{M}_{\Pi}^{(p)},$$

where for a given  $p$  it is used notation  $\mathcal{M}_{\Pi}^{(p)}$  instead of  $\mathcal{M}_{\Pi}$ . Lemma 2.3 is also valid for the corresponding matrices generated by the reference element  $\hat{\mathcal{E}}_0$ , because in their proof it was not important the concrete form of internal functions and the space over them.

**Remark 2.3** Let us suppose that we have orthogonalized the set of the side functions to the set of the internal functions, i.e. instead of  $\mathcal{M}_{II}$  we use  $\mathcal{M}_{(II)}$  such that

$$a(\hat{u}_{II}, \hat{u}_I) = 0, \text{ for any } \hat{u}_I \in \mathcal{M}_I \text{ or for any } \hat{u}_I \in \mathcal{M}_{I,p}, \quad (2.15)$$

for any  $\hat{u}_{(II)} \in \mathcal{M}_{(II)}$ . Then

$$a(\hat{u}_{II}, \hat{u}_{II}) \leq c(1 + \log p)a(\hat{u}, \hat{u})$$

Indeed, taking into account (2.11), (2.15), we can write

$$|\hat{u}_{II}|_{1,\Pi}^2 \leq |\hat{u}_{II} + \hat{u}_I|_{1,\Pi}^2 = |\hat{u} - \hat{u}_{III}|_{1,\Pi}^2 \leq 2(|\hat{u}|_{1,\Pi}^2 + |\hat{u}_{III}|_{1,\Pi}^2) \leq c(1 + \log p)|\hat{u}|_{1,\Pi}.$$

Instead of (2.2) we get

$$\frac{c}{\log p} \sum_{L=I,II,III} a(\hat{u}_L, \hat{u}_L) \leq a(\hat{u}, \hat{u}) \leq 3 \sum_{L=I,II,III} a(\hat{u}_L, \hat{u}_L).$$

## References

1. Babuska I., Craig A., Mandel J., Pitkaranta J. *Efficient preconditioning for the p-version finite element method in two dimensions.* SIAM J. Numer. Anal., Vol. 28, (1991), no.3, pp. 624–661.
2. Pavarino L.F. *Additive Schwartz methods for the p-version finite element method.* Numer. Math. Vol. 66 (1994), no.4, pp. 493–515.
3. Mandel J. *Iterative solvers by substructuring for the p-version finite element method.* Comput. Methods of Appl. Mech. Engrg. Vol. 80 (1990), pp. 117–128.
4. Carey G.F., Barragy E. *Basis function selection and preconditioning high degree finite element and spectral methods.* BIT Vol. 29 (1989), no.4, pp. 784–804.
5. Fisher P.F. *Spectral element solution for the Navier–Stokes equations on high performance distributed memory parallel processors,* in Parallel and Vector Computations in Heat Transfer. HTD–Vol. 133. Ed. J.D.Georgiadis and J.Y.Murphy, 1990.
6. Amon C.N. *Spectral element – Fourier approximation for the Navier –Stokes equations: some aspects of parallel implementation,* in Parallel and Vector computations in Heat Trasfer. HTD Vol. 133. Ed. J.D.Georgiadis and J.Y.Murphy, 1990.
7. George Liu J. W.–H. *Computer solution of large sparse positive definite systems.* Prentice – Hall Inc. Englwood Cliffs. New Jersew 1981.

8. Bahlmann D., Korneev V.G. *Comparison of high order accuracy finite element methods: nodal points selection and preconditioning*. Preprint Nr.237/7.–Chemnitz, Technische Universität Chemnitz, Jg., 1993.
9. Korneev V.G. *On the relation between finite differences and quadratures from Green function for the ordinary second order differential operator*. Vestnik Leningradskogo Universiteta (1972), no.7, pp.36–44 (in Russian).
10. Korneev V.G. *On the construction of the variational–difference schemes of the high orders of accuracy*. Vestnik Leningradskogo Universiteta (1970), no.19, pp.28–40 (in Russian).
11. Korneev V.G. *Finite element methods of the high orders of accuracy* Izdatel'stvo Leningradskogo Universiteta 1970, 205p. (in Russian).
12. Korneev V.G. *Iterative methods for solution of finite element algebraic equations systems*. Zhurnal vychislitel'noy matematiki i matematicheskoi fiziki (1977),V.17, no. 5, pp.1213–1233 (in Russian).
13. Ivanov S.A. and Korneev V.G. *Selection of the high order coordinate functions and preconditioning in the frame of the domain decomposition method* Izvestiya Vyshikh Uchebnykh Zavedeniy (1995),no.4 (395), pp.62–81 (in Russian).
14. Nepomnyaschikh S.V. *Method of splitting into subspaces for solving elliptic boundary value problems in complex–form domains*. Sov. J.Numer. Anal.Math. Modelling (1991). V.6, no. (2), pp. 151–168.